

Order Optimal Coded Delivery and Caching: Multiple Groupcast Index Coding

Mingyue Ji*, Antonia M. Tulino[†], Jaime Llorca[†] and Giuseppe Caire*

* EE Department, University of Southern California. Email: {mingyuej, caire}@usc.edu

[†] Alcatel Lucent, Bell Labs, Holmdel, NJ, USA. Email: {a.tulino, jaime.llerca}@alcatel-lucent.com

Abstract—The capacity of caching networks has received considerable attention in the past few years. A particularly studied setting is the case of a single server (e.g., a base station) and multiple users, each of which caches segments of files in a finite library. Each user requests one (whole) file in the library and the server sends a common coded multicast message to satisfy all users at once. The problem consists of finding the smallest possible codeword length to satisfy such requests. In this paper we consider the generalization to the case where each user places $L \geq 1$ requests. The obvious naive scheme consists of applying L times the order-optimal scheme for a single request, obtaining a linear in L scaling of the multicast codeword length. We propose a new achievable scheme based on multiple groupcast index coding that achieves a significant gain over the naive scheme. Furthermore, through an information theoretic converse we find that the proposed scheme is approximately optimal within a constant factor of (at most) 18.

I. INTRODUCTION

Wireless and mobile data traffic has grown dramatically in the past few years and is expected to increase at an even faster pace in the near future, mainly pushed by on-demand video streaming [1]. Such type of traffic is characterized by *asynchronous content reuse* [2], i.e., the users demands concentrate on a relatively small library of files (e.g., 1000 titles of TV shows and movies), but the streaming sessions happen at arbitrary times such that *naive multicasting* of the same file (e.g., by exploiting the inherent broadcast property of the wireless channel) yields no significant gain. A classical and effective approach to leverage such asynchronous content reuse is *caching* at the user devices. A significant amount of recent work has shown that caching at the wireless edge yields very effective ways to trade-off expensive wireless bandwidth for cheap and widely available storage memory in the user devices, which is the fastest growing and yet untapped network resource in today's wireless networks [3]–[10].

In [3], [4], Llorca *et al.* formulated a general framework of the caching problem in an arbitrary network as an optimization problem and showed the effectiveness of coded delivery via simulations. They further showed the NP-hardness of the problem via an equivalence to network coding in a caching-demand augmented graph. However, due to the generality of the problem, except for a few cases it is unlikely to obtain the optimality of the achievable schemes and the gap between the achievable rate and a converse bound.

In [8], [9], Maddah-Ali and Niesen considered a network formed by a single server (a base station) and n users, where each user place an arbitrary demand to a file in the library and

the server sends a single multicast codeword that is received by all users and satisfies at once all the demands. An achievable scheme based on a combinatorial cache construction (caching phase) and linear coding for the multicast transmission (delivery phase) is proposed in [8] and, through a cut-set information theoretic bound, it is shown that such scheme is near-optimal in the case of arbitrary demands, i.e., in a min-max sense (min codeword length, max over the user demands). This result is extended to the case of random decentralized caching phase in [9].

Consistently with [8], [9], in this paper we refer to the *transmission rate* as the length (expressed in equivalent file units) of the multicast codeword. The order-optimal transmission rate scaling found in [8], [9] is given by $\Theta(\min\{\frac{m}{M}, m - M, n\})$. Notice that in the interesting regime where n and m are large, M is fixed and $nM \gg m$, this yields a multiplicative factor of M in the per-user throughput (in bits per unit time). In passing, we notice here that an alternative caching and delivery approach was proposed in [5], [6], where the communication is also one-hop but takes place between the user nodes (Device-to-Device or “D2D” communications). Both in the random decentralized caching case [5] and in the combinatorial caching and coded delivery case [6], the same rate scaling law is shown to be achievable and approximately optimal for the one-hop D2D network also. In light of these results, it appears that the problem of caching at the user nodes (either in a combinatorial/centralized or random decentralized manner) and delivery (either coded multicast from a single server [8], [9], or uncoded with individual peer-to-peer transmissions [5], or coded in multiple peer-to-multicast transmissions [6]) is well-understood.

In this work, we consider the same setting of [8] (one server, n users, one-hop multicast transmission from the server to the users) where users make multiple requests instead of a single request. This scenario may be motivated by a *FemtoCaching* network [2] formed by n small-cell base stations receiving data from a controlling “macro” base station via the cellular downlink. Each small-cell base station has a local cache of M file units and serves L users through its own local high-rate downlink. Hence, each small-cell base station (which is identified as a “user” in our network model) makes L requests to the macro base station at once. A related relevant work is presented in [11], where an exact solution of the 3-user index coding problem for assigned side information and multiple requests is given. We study the fundamental limits of this

type of network for the general case of n users, m possible messages and L requests per user, where in contrast to the general index coding problem, the side information (i.e., the cache content) at each user is designed as part of the “code” rather than given a priori.

Our contribution is two-fold. First, by using the *same* combinatorial caching phase of [8], we generalize the coded delivery phase by using the directed (fractional) local chromatic number proposed in [12] to the case of multiple groupcasting, where *multiple* means that each user makes $L \geq 1$ requests and *groupcasting* means that one file or sub-packet (see later) may be requested by several users. We show that order gains can be obtained by using the proposed scheme compared to the naive approach of using L times the delivery phase of [8]. Second, we present an information theoretical lower bound of the rate, and show that the proposed scheme meets this lower bound within a constant factor of at most 18.

II. NETWORK MODEL AND PROBLEM FORMULATION

We consider a network with n user nodes $\mathcal{U} = \{1, \dots, n\}$ connected through a single bottleneck link to a server. The server has access to the whole content library $\mathcal{F} = \{1, \dots, m\}$ containing m files of same size of B bits. Each user node has a cache of size M files (i.e., MB bits). The bottleneck link is a shared deterministic channel that transmits one file per unit time, such that all the users can decode the same multicast codeword. At each time unit (slot), each user requests an arbitrary set of L files in \mathcal{F} . Such requests form a matrix \mathbf{F} of size $L \times n$ with columns $\mathbf{f}_u = (f_{u,1}, f_{u,2}, \dots, f_{u,L})^T$ corresponding to the requests of each user $u \in \mathcal{U}$. The caching problem includes two distinct operations: the caching phase and the delivery phase. The caching phase (cache formation) is done a priori, as a function of the files in the library, but does not depend on the request matrix \mathbf{F} . Then, at each time slot, given the current request matrix \mathbf{F} , the server forms a multicast codeword and transmits it over the bottleneck link such that all users can decode their requested files. Formally, we have:

Definition 1: (Cache Phase) The caching phase is a map of the file library \mathcal{F} onto the user caches. Without loss of generality, we represent files as vectors over the binary field \mathbb{F}_2 . For each $u \in \mathcal{U}$, let $\phi_u : \mathbb{F}_2^{mB} \rightarrow \mathbb{F}_2^{MB}$ denote the caching function of user u . Then, the cache content of user u is given by $Z_u \triangleq \phi_u(W_f : f = 1, \dots, m)$, where $W_f \in \mathbb{F}_2^B$ denotes the f -th file in the library. \diamond

Definition 2: (Delivery Phase) A delivery code of rate $R(M, L)$ is defined by an encoding function $\psi : \mathbb{F}_2^{mB} \times \mathcal{F}^{L \times n} \rightarrow \mathbb{F}_2^{R(M, L)B}$ that generates the codeword $X_{\mathbf{F}} = \psi(\{W_f : f = 1, \dots, m\}, \mathbf{F})$ transmitted by the server to the users, and decoding functions $\lambda_u : \mathbb{F}_2^{R(M, L)B} \times \mathbb{F}_2^{MB} \times \mathcal{F}^{L \times n} \rightarrow \mathbb{F}_2^{LB}$ such that each user $u \in \mathcal{U}$ decodes its requested files as $(\hat{W}_{u,f_{u,1}}, \dots, \hat{W}_{u,f_{u,L}}) = \lambda_u(X_{\mathbf{F}}, Z_u, \mathbf{F})$. \diamond

The case of arbitrary demands can be formulated as a *compound channel*, where the delivery phase is designed in order to minimize the rate for the worst-case user demand.

The relevant worst-case error probability is defined as

$$P_e = \max_{\mathbf{F}} \max_{u \in \mathcal{U}} \max_{\ell=1, \dots, L} \mathbb{P}(\hat{W}_{u,f_{u,\ell}} \neq W_{f_{u,\ell}}). \quad (1)$$

The cache-rate pair $(M, R(M, L))$ is achievable if there exist a sequence of codes $\{\phi, \psi, \{\lambda_u : u \in \mathcal{U}\}\}$ for increasing file size B such that $\lim_{B \rightarrow \infty} P_e = 0$. In this context, the system capacity $R^*(M, L)$ (best possible achievable rate) is given by the infimum of all $R(M, L)$ such that $(M, R(M, L))$ is achievable.

III. ACHIEVABLE SCHEME

In this section, we present the proposed achievable scheme. Since arbitrary demands are considered, we simply use the same *sub-packetized caching function* defined in [8], which is optimal (within a fixed factor) for the case $L = 1$. For the sake of completeness, we describe the caching scheme in the following. Let $t = \frac{Mn}{m}$ be a positive integer, then we consider the set \mathcal{P} of all subsets \mathcal{T} (combinations) of distinct users of size t . Each file is divided into $\binom{n}{t}$ sub-packets. For each file, we use $\mathcal{T} \in \mathcal{P}$ to label all the file sub-packets. Then, user u will cache the sub-packets whose label \mathcal{T} contains u . For example, if we have $m = n = 3$ and $M = 1$, denoting users as 1, 2, 3 and files as A, B, C , we have $t = \frac{1 \cdot 3}{3} = 1$ and $\mathcal{P} = \{\{1\}, \{2\}, \{3\}\}$. Each file is divide into $\binom{3}{1} = 3$ sub-packets, i.e., $A = \{A_1, A_2, A_3\}$, $B = \{B_1, B_2, B_3\}$ and $C = \{C_1, C_2, C_3\}$. Then, the user caches are given by

$$Z_1 = \{A_1, B_1, C_1\}, Z_2 = \{A_2, B_2, C_2\}, Z_3 = \{A_3, B_3, C_3\}. \quad (2)$$

Since each user has *part* of each file, the delivery phase consists of providing to each user the *missing part* of the requested files, i.e., the missing sub-packets. For instance, given the caching placement of the example above, if user 1 requests file A , then it needs to obtain from the delivery phase sub-packets A_2 and A_3 . In the following, we denote by \mathcal{S} the set of all implicitly requested sub-packets.

Based on this caching scheme, we design a delivery scheme based on linear index coding (i.e., ϕ is linear function over an extension field of \mathbb{F}_2). In particular, we focus on encoding functions of the following form: for a request matrix \mathbf{F} , the multicast codeword is given by

$$X_{\mathbf{F}} = \sum_{s \in \mathcal{S}} \omega_s \mathbf{v}_s = \mathbf{V} \boldsymbol{\omega}, \quad (3)$$

where ω_s is the binary vector corresponding to sub-packet s , represented as a (scalar) symbol of the extension field \mathbb{F}_κ with $\kappa = 2^{B/\binom{n}{t}}$, $\mathbf{v}_s \in \mathbb{F}_\kappa^\nu$ is the coding vector of sub-packet s and where we let $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_{|\mathcal{S}|}]$ and $\boldsymbol{\omega} = [\omega_1, \dots, \omega_{|\mathcal{S}|}]^T$. The number of rows ν of \mathbf{V} yields the number of sub-packet transmissions. Hence, the coding rate is given by $R(M, L) = \nu / \binom{n}{t}$ file units.

In order to design the coding matrix \mathbf{V} for multiple group-casting, we shall use the method based on directed local chromatic number introduced in [12]. The definition of directed local chromatic number is given as follows:

Definition 3: (Directed Local Chromatic Number (I)) The directed local chromatic number of a directed graph \mathcal{H}^d is defined as:

$$\chi_l(\mathcal{H}^d) = \min_{c \in \mathcal{C}} \max_{o \in \mathcal{O}} |c(\mathcal{N}^+(o))| \quad (4)$$

where \mathcal{C} denotes the set of all vertex-colorings of \mathcal{H} , the undirected version of \mathcal{H}^d , \mathcal{O} denotes the vertices of \mathcal{H}^d , $\mathcal{N}^+(o)$ is the closed out-neighborhood of vertex o ,¹ and $c(\mathcal{N}^+(o))$ is the total number of colors in $\mathcal{N}^+(o)$ for the given coloring c . \diamond

An equivalent definition of the directed local chromatic number in terms of an optimization problem is given by:

Definition 4: (Directed Local Chromatic Number (II)) Let \mathcal{I} denote the set of all independent sets of \mathcal{H} , the directed local chromatic number of \mathcal{H}^d is given by:

$$\begin{aligned} & \text{minimize} \quad k \\ & \text{subject to} \quad \sum_{I: \mathcal{N}^+(o) \cap I \neq \emptyset} x_I \leq k, \quad \forall o \in \mathcal{O} \end{aligned} \quad (5)$$

$$\sum_{I: o \in I} x_I \geq 1, \quad \forall o \in \mathcal{O} \quad (6)$$

$$x_I \in \{0, 1\}, \quad \forall I \in \mathcal{I} \quad (7)$$

\diamond

Also, we have:

Definition 5: (Directed Fractional Local Chromatic Number) The directed fractional chromatic number is given by (5) when relaxing (7) to $x_I \in [0, 1]$. \diamond

In the following, we describe the proposed scheme when t is a positive integer. When t is not an integer, we can simply use the resource sharing scheme for caching proposed in [8] and achieve convex combinations of the cases where t is an integer.

In order to find the coding matrix \mathbf{V} we proceed in three steps [12]: 1) constructing the directed conflict graph \mathcal{H}^d ; 2) computing the directed local chromatic number $\chi_l(\mathcal{H}^d)$ and the corresponding vertex-coloring c ; 3) constructing \mathbf{V} by using the columns of the parity check matrix of a $(|c|, \chi_l(\mathcal{H}^d))$ -MDS code.² The detailed delivery scheme is described in the following.

1) Construction of the directed conflict graph \mathcal{H}^d .

- Consider each sub-packet requested by a user as a *distinct* vertex in \mathcal{H}^d . This means that if a certain sub-packet is requested N times, for some N , because it appears in multiple requests from different users, it corresponds to N different vertices in \mathcal{H}^d . Since each user caches $\binom{n-1}{t-1}$ sub-packets of each file, and each file is partitioned into $\binom{n}{t}$ sub-packets, the number of requested sub-packets of each file for each user is $\binom{n}{t} - \binom{n-1}{t-1} = \binom{n-1}{t}$

¹Closed out-neighborhood of vertex o includes vertex o and all the connected vertices via out-going edges of o .

²According to the classical coding theory notation, an (ρ, ν) -MDS code is a code with length ρ , dimension $\rho - \nu$, and minimum Hamming distance $d = \nu + 1$. An MDS code is able to correct up to ν erasures. This implies that a linear MDS code has parity-check matrix (of dimensions $\nu \times \rho$ such that any subset of $\ell \leq \nu$ columns form a $\nu \times \ell$ submatrix of rank ℓ .

(by Pascal's triangle). Hence, the total number of files requested Ln corresponds to $|\mathcal{O}| = Ln \binom{n-1}{t}$ vertices in \mathcal{H}^d .

- For any pair of vertices o_1, o_2 , we say that vertex (sub-packet) o_1 interferes with vertex o_2 if o_1 is not in the cache of the user(s) who requests o_2 , and o_1 and o_2 do not represent the same sub-packet. Then, draw a directed edge from vertex o_2 to vertex o_1 if o_1 interferes with o_2 .

2) Color the directed conflict graph \mathcal{H}^d according to the coloring which gives the directed local chromatic number $\chi_l(\mathcal{H}^d)$. The total number of colors needed to give the directed local chromatic number is denoted by $|c|$.

3) Construct \mathbf{V}' be the parity check matrix of an MDS code with parameters $(|c|, \chi_l(\mathcal{H}^d))$ over \mathbb{F}_κ . Notice that it is well-known that for sufficiently large κ such MDS code exists. Since we are interested in $B \rightarrow \infty$ and $\kappa = 2^{B/\binom{n}{t}}$, for sufficiently large B this code exists.

4) Allocate the same coding vector to all the vertices (sub-packets) with the same color in c . Then, \mathbf{V} is obtained from the MDS parity-check matrix \mathbf{V}' by replication of the columns, such that each column in \mathbf{V} is replicated for all vertices with the same corresponding color in c . Eventually, all packets are encoded using the linear operation in (3).

The above constructive coding scheme proves the following (achievable) upper bound on the optimal coding length:

Theorem 1: The optimal coding length for the multiple group cast problem with integer $t = nM/m$ satisfies

$$R^*(M, L) \leq R^{\text{LC}}(M, L) = \max_{\mathbf{F}} \frac{\chi_l(\mathcal{H}^d)}{\binom{n}{t}}. \quad (8)$$

where the upper bound is achieved by the caching and linear coded delivery scheme seen above. \square

To illustrate our proposed scheme, we consider the following example.

Example 1: Let $m = n = 3$, $M = 1$ and $L = 2$. We denote the files as A, B, C as before. The caching placement is given by (2). Let user 1 request file A, B , user 2 request file B, C and user 3 request file C, A . In this case, user 1 requests sub-packets A_2, A_3, B_2, B_3 . User 2 requests sub-packets B_1, B_3, C_1, C_3 . User 3 requests sub-packets C_1, C_2, A_1, A_2 . The resulting conflict graph \mathcal{H}^d and the corresponding vertex-coloring is shown in Fig. 1, where we can see that, in total, 6 colors are used but locally each vertex needs at most 5 colors, such that $\chi_l(\mathcal{H}^d) = 5$. A parity check matrix \mathbf{V}' of a $(6, 5)$ MDS code is given by

$$\mathbf{V}' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \quad (9)$$

Then, we allocate the same vector to the vertex (sub-packet)

with the same color. We obtain

$$\begin{aligned} A_1, A_3 : & \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} & A_2 : & \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} & B_1, B_2 : & \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \\ C_2, C_3 : & \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} & B_3 : & \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} & C_1 : & \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned} \quad (10)$$

The transmitted codeword is given by $A_1 \oplus A_3 \oplus C_1$, $A_2 \oplus C_1$, $B_1 \oplus B_2 \oplus C_1$, $C_2 \oplus C_3 \oplus C_1$, $B_3 \oplus C_1$, of length $5/3$ file units. \diamond

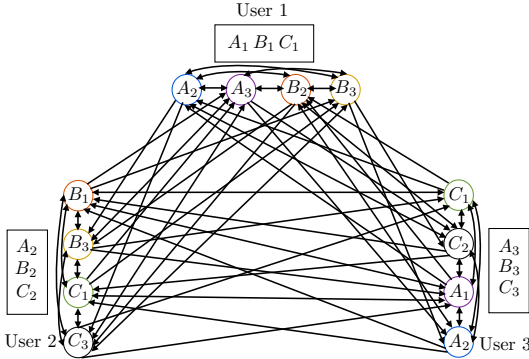


Fig. 1. An example for the proposed caching and coding scheme, where $n = m = 3$, $M = 1$. Files are denoted by A, B, C . We let user 1 request A, B , user 2 request B, C , and user 3 request C, A . Each file is partitioned into 3 sub-packets. The rectangular represents the cached sub-packets. The vertices represent the requested sub-packets for each user. The graph is the conflict graph \mathcal{H}^d constructed by the proposed algorithm. We assign a color to each vertex, which is shown on the perimeter of each vertex.

We can also consider a delivery scheme based on the directed fractional local chromatic number, as given in Definition 5, that achieves a generally better performance (see [13] for details). For the optimality of index coding based on local chromatic number, where only the index coding delivery phase is considered for assigned node side information (i.e., without optimizing caching functions), it can be shown that the gap between the proposed index code and the converse is bounded by the integrality gap of a certain linear programming (see [13] for details).

IV. PERFORMANCE ANALYSIS

To show the effectiveness of the proposed coded caching and delivery scheme, we first consider the two special points of the achievable rate versus L corresponding to $L = 1$, considered in [8], and $L = m$, i.e., when every user requests the whole library.

Theorem 2: When each user makes only one request ($L = 1$), then the achievable rate of the proposed scheme satisfies

$$R^{LC}(M, 1) \leq R^{MN}(M), \quad (11)$$

where $R^{MN}(M)$ is the convex lower envelope of $\min \left\{ n \left(1 - \frac{M}{m} \right) \frac{1}{1 + \frac{Mn}{m}}, m - M \right\}$, achieved by the scheme of [8]. \square

Theorem 3: When each user requests the whole library ($L = m$), then the achievable rate of the proposed scheme satisfies

$$R^{LC}(M, m) \leq m - M. \quad (12)$$

\square

The rate $m - M$ can also be achieved, with high probability for a large field size, by using random linear coding. Moreover, it can be shown that this is information theoretically optimal.

Theorems 2 and 3 show that in the extreme regimes of L the proposed scheme performs at least as good as the state of the art scheme in literature. A general upper bound of the achievable rate is given by:

Theorem 4: The achievable rate of the proposed scheme satisfies

$$R^{LC}(M, L) \leq R_{ub}^{LC}(M, L) \quad (13)$$

where $R_{ub}^{LC}(M, L)$ is the convex envelope (with respect to M) of

$$\min \left\{ Ln \left(1 - \frac{M}{m} \right) \frac{1}{1 + \frac{Mn}{m}}, m - M \right\}.$$

\square

The qualitative performance of the proposed scheme is shown in Fig. 2(a), where we call the scheme that repeats the delivery scheme designed for one request (in [8]) L times as *direct scheme*. To measure the performance of the proposed scheme quantitatively, we let $L = \alpha m$, where $\alpha \in (0, 1)$. Then, let M be a constant, as $m \rightarrow \infty$, we can see that the rate of the direct scheme is $L \cdot R^{MN}$, which scales as $\Theta(m^2)$. While by using (13), the rate of the proposed scheme scales at most $O(m)$. Thus, we can see the proposed scheme can have an order gain compared to the direct scheme.

In order to appreciate the gain achieved by the proposed scheme over the direct scheme for fixed parameters, we consider the case of $n = m = 3$ and $M = 1$, as in Example 1. In this case, though, the requests can be arbitrary. Fig. 2(b), shows the worst-case (over the requests) coding rate versus L and we can observe that, even for small n, m and M , the gain achieved by the proposed scheme with respect to the direct scheme is fairly large. For example, for $L = 2$, the proposed scheme requires $\frac{5}{3}$ file units to satisfy any request, while the directed scheme or random linear coding require 2 file units.

V. CONVERSE AND OPTIMALITY

To show the optimality (up to a constant factor) of the proposed scheme, we prove an information theoretic lower bound on the rate by using the cut-set bound technique:

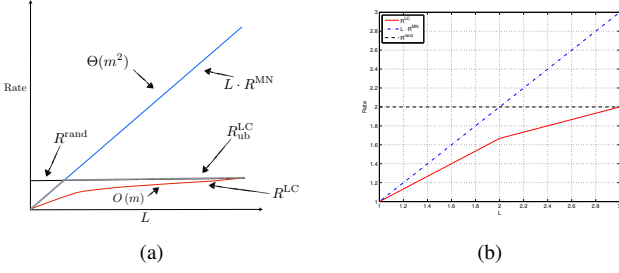


Fig. 2. (a). A qualitative illustration of the different delivery schemes (M is assumed to be not very small). $L \cdot R^{\text{MN}}$ (blue curve) represents rate by using the scheme in [8] L times. R^{rand} (black curve) is the rate by random linear coding. $R^{\text{LC}_{\text{ub}}}$ (grey curve) is an upper bound of the achievable rate of the proposed scheme. R^{LC} (red curve) represents the rate by the proposed scheme based on directed local chromatic number. (b). An example of the rate by the proposed scheme. In this example, $n = m = 3$ and $M = 1$. In this figure, all the symbols have the same meanings as in Fig. 2(a).

Theorem 5: The optimal coding length for the multiple group cast problem satisfies

$$R^*(M, L) \geq R^{\text{lb}}(M, L) = \max \left\{ \max_{s=1, \dots, \min\{\lfloor \frac{m}{L} \rfloor, n\}} \left(Ls - \frac{sM}{\lfloor \frac{m}{L} \rfloor} \right), \max_{s=1, \dots, \min\{\lfloor \frac{m}{L} \rfloor, n\}} (Ls - sM), \frac{m-M}{\lceil \frac{m}{L} \rceil} \right\}. \quad (14)$$

□

In fact, we can show that the achievable upper bound of Theorem 1 and the converse lower bound of Theorem 14 have a bounded ratio:

Theorem 6: The multiplicative gap between the upper and lower bounds satisfies

$$\frac{R^{\text{LC}}(M, L)}{R^{\text{lb}}(M, L)} \leq 18. \quad (15)$$

□

From simulation, we observed that this multiplicative gap is generally smaller than 5. For example, when $m = n = 100$ and $M = 20$, we found a gap always less than 3.88.

VI. DISCUSSIONS

For the index coding problem, since the side information is assigned, it is generally difficult to find a constant multiplicative gap between the achievable rate and the converse. For example, in [14], the authors consider a random unicast index coding problem, where each user can cache βm files uniformly and $\beta \in (0, 1)$. It is assumed that $m = n$ and all the users request different files. In this case, the best *multiplicative* gap between the achievable rate and the converse lower bound is $\Theta\left(\frac{\sqrt{n}}{\log n}\right)$, which goes to infinity as $n \rightarrow \infty$. However, for the caching problem, due to the fact that we can design the caching functions, it is possible to find coding schemes with a bounded multiplicative gap for arbitrary (worst-case) demands. In the

case treated in this paper, for example, any scheme (possibly non-linear) would provide at most a factor of $18 \times$ gain with respect to the proposed linear scheme.

We notice from Section IV that, to achieve the order gain compared with the direct scheme, we need to code over all the L requested files simultaneously, in contrast with the repeated application of the scheme in [8] (direct scheme), where the user can decode instantaneously. Therefore, the proposed scheme may be useful in the FemtoCaching application, where each user in our system corresponds to a local server (a small-cell base station) serving L requests to its own local users on a much faster local connection. In this case, there is no natural ordering of the L requests such that there is no interest to decode them instantaneously.

REFERENCES

- [1] Cisco, "The Zettabyte Era-Trends and Analysis," 2013.
- [2] N. Golrezaei, A.F. Molisch, A.G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *Communications Magazine, IEEE*, vol. 51, no. 4, pp. 142–149, 2013.
- [3] J. Llorca, A.M. Tulino, K. Guan, and D.C. Kilper, "Network-coded caching-aided multicast for efficient content delivery," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3557–3562.
- [4] J. Llorca and A.M. Tulino, "The content distribution problem and its complexity classification," Alcatel-Lucent technical report, 2013.
- [5] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *arXiv:1312.2637*, 2013.
- [6] M. Ji, G. Caire, and A.F. Molisch, "Fundamental limits of distributed caching in d2d wireless networks," *arXiv:1304.5856*, 2013.
- [7] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv:1305.5216*, 2013.
- [8] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *arXiv:1209.5807*, 2012.
- [9] M. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," *arXiv:1301.5848*, 2013.
- [10] S. Gkitzenis, GS Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *arXiv:1201.3095*, 2012.
- [11] S. Unal and A.B. Wagner, "General index coding with side information: Three decoder case," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2013, pp. 1137–1141.
- [12] K. Shanmugam, A. G. Dimakis, and M. Langberg, "Local graph coloring and index coding," *arXiv:1301.5359*, 2013.
- [13] M. Ji, A.M. Tulino, J. Llorca, and G. Caire, "Coded caching for efficient content delivery: Multiple requests, multiple group-cast index coding," *In Preparation*.
- [14] I. Haviv and M. Langberg, "On linear index coding for random graphs," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2231–2235.